DOCUMENT RESUME

ED 441 823 TM 030 858

AUTHOR Nering, Michael L.; Bay, Luz G.; Meijer, Rob R.
TITLE Validity of Student Scores in the New Hampshire Educational Assessment and Improvement Program.
PUB DATE 2000-04-00
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Responses; *Scores; State Programs; *Testing Problems; Testing Programs; *Validity
IDENTIFIERS Low Stakes Tests; New Hampshire; *New Hampshire Educational Improve and Assess Prog; *Repeating Response Tendency

ABSTRACT

In state assessment programs in which performance has no real immediate consequence for the individual examinee, the issue of examinee motivation arises. Some examinees may respond to questions in ways that do not reflect their real knowledge of the test domain. In this study, a new approach was developed to identify students who have responded to questions in a way that does not reflect their knowledge adequately. The new method was designed to detect examinees that responded according to some repeating pattern. The new method was then compared to two indices previously developed by Drasgow, Levine, and Williams (1985) using data from New Hampshire's statewide mathematics assessment for 16,630, 16,387, and 14,445 students from grades 3, 6, and 10 respectively. These indices were found to be distributed in a manner that was more or less expected, something that suggests that the indices may be useful in detecting examinees that are responding in a way not in accord with the underlying test model. However, results suggest that these indices will only occasionally detect responding according to a repeating pattern. Use of the new method, called the PM method, to detect responding in a repeated pattern is supported by the study. (Contains 8 tables and 23 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 823

VALIDITY OF STUDENT SCORES IN THE NEW HAMPSHIRE EDUCATIONAL ASSESSMENT AND IMPROVEMENT PROGRAM

MICHAEL L. NERING
LUZ G. BAY
ADVANCED SYSTEMS IN MEASUREMENT AND EVALUATION, INC.

ROB R. MEIJER
UNIVERSITY OF TWENTE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

M. Nering

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 2000. All data used in this investigation has been used by permission of the New Hampshire Department of Education. Correspondence concerning this paper should be sent to: Michael L. Nering, Advanced Systems, 171 Watson Road, Dover, NH 03820, USA. Email: MNering@asme.com

TM030858

BEST COPY AVAILABLE

Validity of Student Scores in the New Hampshire Educational Assessment and Improvement Program

In virtually every state in the United States various assessment programs are used to evaluate the quality of education based on the curriculum and the degree to which teachers adhere to a curriculum. Many of these assessment programs measure examinee performance on a test, but results are not primarily used to evaluate the individual examinee. The interest is mainly at the aggregate level of the school, district, or state. In some states student performance on a statewide assessment are used to make decisions regarding various policies and programs, to evaluate programs for teachers such as in-service training (Darling-Hammond, 2000), and for school accountability and personnel evaluation (Dorn, 1998). In state assessment programs where performance has no real immediate consequence to the individual examinee, motivation becomes a factor that threatens the validity of inferences made based on student scores. This paper will focus on examinees who may not have been motivated when taking the test and in effect may have responded to test question in ways that do not reflect their true knowledge of the test domain. In an assessment composed of mainly multiple-choice items, the lack of motivation may produce responses composed of repeating patterns. The degree to which this practice exists reflects the validity, or the lack thereof, of decisions made based on student performance on state assessment.

The issue of examinee motivation is not new in educational and psychological measurement. Measurement specialists have tried to find the effects of examinee motivation on test results (e.g., Wolf, 1995), or have tried to identify unmotivated examinees through person-fit analysis where the observed item scores are compared with the expected scores on the basis of some test model (e.g., Birenbaum, 1986). As noted by several authors (e.g. Klauer, 1991, Meijer & Sijtsma, in press) a drawback of person-fit procedures is that the rate to detect specific types of misfitting patterns is low. Moreover, only deviations against the model are tested which may result in interpretation problems. For example, misfitting item score patterns may be the result of a number of causes such as cheating behavior, misunderstanding of the questions, or even clerical errors. Several authors therefore have proposed to test against specific model violations (Drasgow , Levine, & McLaughlin, 1987) or to study local violations in an item score pattern (e.g., Sijtsma & Meijer, in press). Because we are interested in the detection of unmotivated examinees we will focus on a specific type of misfitting behavior, producing repeated pattern of item responses, which is often described as the result of unmotivated test response behavior (e.g., Haladyna, 1994, p 165; Schmitt et al. 1999). For example, suppose a test consists of multiple choice items where for each item there exists four alternatives A, B, C, D. Assuming that the correct answers are randomly distributed across the items as is often the case, examinees responding say ABCDABCD… are clearly not taking the test seriously and should not be included in results that are used to determine the ability of the student which are in turn used to evaluate the quality of the education that the student is receiving. A pattern based index will assist in identifying students that may not have been motivated during the assessment administration. Identifying these examinees prior to item calibration, equating, and score reporting may help to improve the usefulness of results from a large scale assessment program.

In reporting scores to an examinee the New Hampshire Educational Improvement and Assessment Program (NHEIAP) assessment offers scale scores, proficiency levels, along with various descriptions of how the student has performed. There is also a "validity index" that is used to indicate whether or not the score that is reported is an

BEST COPY AVAILABLE

accurate representation of the student's ability level. In describing the validity index on the student report there is a statement claiming that the index is used to reflect "lack of effort, not feeling well on one or more days of testing, or other similar considerations." This index is index is also used to identify an examinee that may not be taking the test seriously. For example, an examinee may respond ABCDCBA.... or they may respond ABABABAB... because it offers a response pattern in a test booklet that is visually appealing. In the NHEIAP assessment examinees are identified as having invalid scores if they have elicited a repeating pattern to a consecutive set of items belonging to at least half of the test. For example, if an examinee responded say ABCDABCD ... and so forth to the first 16 items on a 32 item assessment then they would be identified as "invalid". Response strings that are considered suspicious are: all one response alternative (e.g., all As), ABCD, DCBA, ABAB, BABA, CDCD, DCDC, and so forth.

There could exist a large number of other types of patterns that students may use that could be due to lack of motivation. Obviously, these sorts of response patterns do not represent student ability and the validity index used in the NHEIAP assessment is one way that these students can be identified. The main limitation of the NHEIAP validity index is that it many of the decisions that are used (i.e., a repeating pattern on at least half of the test and the specific patterns that are searched for) are arbitrary in nature, and it is limited by the number of patterns that the measurement specialist can think of ahead of time. Clearly, there is a need for a statistical approach that reflects the likelihood that a repeating-strings pattern would result.

The aim of this study is to propose a new fit statistic that is sensitive to item score patterns with repeating strings of item scores. The detection rate of this new statistic is compared with existing methods to detect misfitting item score patterns.

Previous Research

Item response theory (IRT) models describe the probability of a correct response to an item as a function of the item and person parameters (e.g., Hambleton & Swaminathan, 1985, pp. 35-48). Both IRT models for dichotomous item scores and models for polytomous item scores have been proposed. In this paper we will focus on dichotomous items. An often used unidimensional IRT model is the three-parameter logistic model which can be defined as:

$$P_j(\hat{\theta}) = c_j + (1 - c_j)\frac{\exp\left[Da_j(\hat{\theta} - b_j)\right]}{1 + \exp\left[Da_j(\hat{\theta} - b_j)\right]} \qquad (1)$$

where: j indexes the item,

a represents the item discrimination or slope parameter,

b represents the item difficulty,

c represents the lower asymptote (pseudo-guessing) parameter,

$\hat{\theta}$ represents an ability estimate for an examinee, and

D represents the normalizing constant (1.701).

The IRT model in Equation 1 reduces to the two-parameter model when c=0 for all items, and the one-parameter or Rasch model when both c=0 and a=1 for all items. To investigate the fit of an item score pattern to an IRT model, most person-fit statistics have been proposed that are designed to investigate the likelihood of an item score pattern under the null hypothesis of fitting response behavior. As discussed in Meijer and Sijtsma (1999) most person-fit statistics can be expressed in a similar manner. For example, if we let X represent an item score (0/1) to item j (where j=1,2,3, ..., J) and $w_j(\hat{\theta})$ represents a suitable function then a general form in which most person-fit statistics using dichotomous item scores can be expressed as:

$$U = \sum_{j=1}^{J} \left[X_j - P_j(\hat{\theta}) \right] w_j(\hat{\theta}) \tag{2}$$

U can then used as an index to indicate the extent to which an specific response pattern is in agreement with the test model.

Many researchers have evaluated several different methods of determining examinee-model fit, and the most promising tool in person-fit research has been the l_z index proposed by Drasgow, Levine, and Williams (1985). The l_z index reflects the standardized version of the ordinate of the likelihood function for a particular response pattern. The ordinate of the likelihood function can be determined for particular response pattern by:

$$l_o = \sum_{j=1}^{J} X_j \ln P_j(\hat{\theta}) + (1 - X_j) \ln \left[1 - P_j(\hat{\theta}) \right]. \tag{3}$$

The standardized version of Equation 1 can be found by simply subtracting the expected value of l_o and dividing by the square root of the variance of l_o for a given ability level. Thus,

$$l_z = \frac{l - E(l_o)}{\mathrm{var}(l_o)^{1/2}}. \tag{4}$$

The expected value and variance terms can be respectively defined as:

$$E(l_o) = \sum_{j=1}^{J} \left\{ P_j(\hat{\theta}) \ln P_j(\hat{\theta}) + \left[1 - P_j(\hat{\theta}) \right] \ln \left[1 - P_j(\hat{\theta}) \right] \right\}, \text{ and} \tag{5}$$

$$\text{var}(l_o) = \sum_{j=1}^{J} P_j(\hat{\theta})\left[1 - P_j(\hat{\theta})\right]\left\{\ln \frac{P_j(\hat{\theta})}{1 - P_j(\hat{\theta})}\right\}^2 . \qquad (6)$$

The use of l_z has been investigated under a variety of experimental conditions and compared to several other person-fit statistics (e.g., Drasgow, Levine, & McLaughlin, 1987, 1991). However, the use of l_z index is restricted to dichotomously scored data (0/1) rather than the unscored (ABCD) data. Thus, when using traditional methods of indexing person fit a great deal of information pertaining to how the examinee has responded to items in the test booklet is not used.

Fortunately, Drasgow, Levine, and Williams (1985) also derived the l_{zh} index that can be used to determine examinee model fit when polytomous data are used instead of the dichotomous data. Although the l_{zh} index has received little attention by person fit research, it may be a useful tool in statewide assessments where examinees may not be motivated during the test administration. Drasgow et al. developed this index in an attempt to use all the information associated with multiple choice items. They did not define the l_{zh} index under a specific IRT model, but instead developed an index that was not dependent on some underlying test model[1]. In presenting the l_{zh} index we first begin by defining

$$V = \{V_1, V_2, \ldots, V_J\} \qquad (7)$$

as representing a random vector of response options to the set of J items. A specific observed vector of response options can defined as:

$$v = \{v_1, v_2, \ldots, v_J\}. \qquad (8)$$

The l_{zh} index can then be defined as:

$$l_{zh} = \sum_{j=1}^{J}\sum_{m=1}^{A} \delta_m(v_j) \ln P_{jm}(\hat{\theta}) \qquad (9)$$

where m indexes the response alternatives (m=1,2, ... A),

P_{jm} represents the proportion of examinees responding to alternative m of item j,

$\delta_m(v_j)$ is equal to 1 if the observed response is equal to m, otherwise 0, and

[1] It may be more appropriate to say that l_{zh} is less model dependent than l_z. In Equation 6 we see that we still use an estimate of an ability value, and this is determined through typical ability estimation procedures using IRT. However; the calculation of the index is done using the raw data rather than directly from the likelihood function.

all other terms have been previously defined.

Just as with the l_z index the l_{z^h} index is found by subtracting the expected value and variance from l_{oh}:

$$l_{z^h} = \frac{l_{oh} - E(l_{oh})}{\mathrm{var}(l_{oh})^{1/2}}. \tag{10}$$

The expected values and variance terms that can be respectively defined as:

$$E(l_{oh}) = E\left\{\sum_{j=1}^{J}\sum_{m=1}^{A}\delta_m(V_j)\ln P_{jm}(\hat{\theta})\right\} = \sum_{j=1}^{J}\sum_{m=1}^{A}P_{jm}(\hat{\theta})\ln P_{jm}(\hat{\theta}), \text{ and} \tag{11}$$

$$\mathrm{var}(l_{oh}) = \mathrm{var}\left\{\sum_{j=1}^{J}\sum_{m=1}^{A}\delta_m(V_j)\ln P_{jm}(\hat{\theta})\right\} = \sum_{j=1}^{J}\left[\sum_{m=1}^{A}\sum_{k=1}^{A}P_{jm}(\hat{\theta})P_{jk}(\hat{\theta})\ln P_{jm}(\hat{\theta})\ln\left(\frac{P_{jm}(\hat{\theta})}{P_{jk}(\hat{\theta})}\right)\right] \tag{12}$$

where k is similar to m and indexes the response options. Van Krimpen-Stoop and Meijer (2000) compared the empirical distribution of l_{z^h} with the standard normal distribution for both paper -and-pencil tests and computer adaptive tests, using the partial credit model (Masters, 1992). The results showed that although the mean and the standard deviation were slightly different from the expected 0 and 1, the empirical type I errors were close to the nominal levels, for both paper-and-pencil tests and adaptive tests.

The main limitation of many previously developed indexes of person fit, including the l_z and l_{z^h} indices, is that they do not give information about the type of misfitting response behavior (Nering & Meijer, 1998). An examinee is simply identified as having a fit index that is relatively large in value, and no information about the response behavior is provided. This is because these statistics do not test against a specific alternative. An alternative approach was, for example, followed by Klauer (1991) who defined a test statistic for the hypothesis that examinee's is invariant over subtests of the total test.

Purpose of Study

In this study, we will first investigate the distribution of l_z and l_{z^h} using statewide assessment data. This will be done by studying the first four moments of the distribution for each statistic. Typically, standardized person-fit statistics are assumed to follow a standard normal distribution (e.g., Drasgow, Levine, & Williams, 1985); however, for dichotomous data, researchers have found that this depends on the test length (Nering, 1995, 1997). For long tests (larger than, say, 60 items) empirical and nominal type I errors are often reasonably in agreement with empirical type I errors. This is because for long tests the latent trait can be estimated precisely, for short tests this is

a problem and as a result the variation of the distribution of a person-fit statistic is reduced (Meijer & Sijtsma, in press). Additional research is needed to study how person-fit statistics are distributed in the context of a low-stakes assessment where examinees may not be taking the test seriously. In addition to the van Krimpen-Stoop & Meijer (2000) paper, research is also needed to determine the usefulness of l_{z^h} using emprirical data and to study whether or not this index follows a standard normal distribution. Given that the l_{z^h} index uses polytomous data instead of dichotomous data like l_z the two statistics may operate in a complimentary manner in determining examinee-model fit. That is, certain types of nonmodel-fitting behaviors may be detected with the l_z index while other types of nonmodel-fitting behaviors may be detected with the l_{z^h} index. Most importantly, using these traditional methods for determining the extent to which examinees are responding in accordance with the underlying test model will be important in understanding and interpreting results from a statewide assessment.

The second purpose of this study is to develop an approach that identifies examinees that responded to test questions by generating a repeated pattern of item responses rather than according to their ability. This study will compare the detection rates of the new approach to the previously developed l_z and l_{z^h} indexes. Differences found between the two approaches will determine which method is best at identifying unmotivated examinees. Because the approaches are fundamentally different it is expected that different students will be identified as responding in an unexpected manner, and that using the approaches together may be the best way of evaluating examinee-model fit. An empirical example will be given where NHEIAP data will be used and examinees that have pattern based response patterns will be identified.

Hypothesis Testing to Determine the Validity of Scores

The fundamental idea behind finding pattern-based responses is that it is reasonable to conjecture that an unmotivated examinee may simply respond to items using some sort of consecutive repeating pattern. For example, an examinee may respond with all As, or ABABAB. There is empirical evidence that this indeed may occur as a result of unmotivated test taking. For example, Freund and Rock (1992) identified examinees, using a visual inspection method, who appeared to be using some type of pattern to mark their answer sheets, as opposed to marking their responses on the basis of the perceived correct answer. Also, Paris, Lawton, Turner, & Roth (1991) discussed the use of pattern marking of school age children.. From this perspective we can think of repeat patterns as single repeats (e.g., all As or all Bs), as double repeats (e.g., CDCD), as triple repeats (CDBCDB), and so forth. Another reasonable conjecture is that after responding to a portion of the test, the student would respond to a subsequent portion by repeating the pattern already made. For example in a 50 item test a student might respond to the first 25 items and then copy those responses to the second half of the test. A hypothesis testing approach will be used to detect examinees who engage in these types of behavior.

Suppose student S took a multiple-choice test with n items. Let S's response vector be represented as V (see Equation 7). If there exists a string of responses of length m, where $m \leq n/2$ or $m \leq (n-1)/2$, and this pattern appears in V more than once, can we infer that S responded to the test questions according to his ability? Or did S respond by repeating a pattern? The null hypothesis is that S responded to all the items based on his ability. In

rejecting the null hypothesis the conclusion would be that *S* responded to some portion of the test by repeating some pattern.

In an *n* item test, the longest string that could be repeated would be half of *n* or half of *n-1*, depending on whether *n* is even or odd. If *n* is even and $m=n/2$, one could compare the response string from the first half of the test to that from the second half of the test. That is comparing the response vector $(v_1, v_2,...,v_m)$ to the response vector $(v_{m+1}, v_{m+2},...,v_n)$. If *n* is odd and $m=(n-1)/2$, the response vector $(v_1, v_2,...,v_m)$ could be compared to two vectors: $(v_{m+1}, v_{m+2},...,v_{n-1})$ and $(v_{m+2}, v_{m+3},...,v_n)$.

In general, the response vector $(v_j, v_{j+1},...,v_{j+r})$may be repeated by an examinee in any of the following vectors: $(v_{j+r+1}, v_{j+r+2},...,v_{j+2r+1})$, $(v_{j+r+2}, v_{j+r+3},...,v_{j+2r+2})$,..., $(v_{n-m}, v_{n-m+1},...,v_{n-1}, v_n)$. That is, any response vector of length $r+1$ within the test of *n* items could be repeated by an examinee in $n-2r-j$ possible subsequent portions of the test, where $r=1, 2,...,(n-j)/2$ or $(n-j+1)/2$, depending on whether $n-j$ is even or odd.

Suppose for some $k=1,2,..., n-2r-j$, the strings $s_0=(v_j, v_{j+1},...,v_{j+r})$ and $s_k=(v_{j+r+k}, v_{j+r+k+1},...,v_{j+2r+k})$ are identical. If s_0 is considered fixed, and assuming local item independence, the probability that $s_0 = s_k$ is the product of probabilities associated with *S*'s responses to the $r-1$ items in s_k. That is,

$$P_k = P(s_k = s_0) = \prod_{i=0}^{r} P(v_{j+r+k+i}). \qquad (13)$$

To make a decision whether or not *S* repeated s_0, one has to establish the risk of inferring that *S* repeated a pattern if in fact he/she responded to the questions according to his ability. This is the probability of type I error, α. To control the probability of making a type I error a *Bonferroni* correction procedure was used (e.g., Dunn, 1961). Using this procedure if α^* is the probability of making the wrong inference in one comparison, then the probability of at least one wrong decision in *x* comparisons is $1-(1-\alpha^*)^x$. Letting $\alpha = 1-(1-\alpha^*)^x$, $\alpha^* = 1-\sqrt[x]{1-\alpha}$, making α^* the type I error rate for each comparison. Because s_0 may be compared with $n-2r-j$ strings, then $x=n-2r-j$. Thus, if $P_k \le \alpha^*$, then we reject the null hypothesis that *S* responded to the multiple-choice questions according to his/her ability.

Because this approach either results in rejecting or accepting the null hypothesis for a student, a dichotomous index, *PM*, can be found for each student. *PM* will equal 1 when the null hypothesis is rejected (indicating that the student response pattern is suspicious), or *PM* will be equal to 0 when the null hypothesis is accepted (indicating nothing unusual about the student response pattern).

Empirical Example

Methods

Dataset Calibration. For this study datasets from the Math portion of the NHEIAP assessment for grades 3, 6, and 10 were used. In each grade there were 8 test forms and the total number of students taking part in the 1998-1999 assessment was 16,630, 16,387, and 13,445 for grade 3, 6, and 10, respectively. Approximately an equal

BEST COPY AVAILABLE

number of students within a given grade took each of the test forms. Each of the tests contained both multiple-choice (MC) items and open-ended (OE) items, and items were either *matrix* or *common*. In the case of the OE items it is difficult for the examinee to control what score he/she will be assigned. Because the examinee has direct control in the pattern created in the response string with the MC items the OE are of less interest and will not be used in this study. Common items were administered to all students, while matrix items were administered to students that had taken a particular form. For example, in grade 3 there were 32 common MC items and 8 MC matrix on each form. Thus, there were a total of 96 MC items on the Math grade 3 test [32 common plus 8 matrix on each of the 8 forms (32+(8*8)=96)]. For grades 6 and 10 there were 24 common items and 10 matrix items per 8 forms for a total of 104 items [24+(8*10)=104].

The program Parscale (1999) was used to fit a two-parameter logistic IRT model to the datasets. Separate calibrations were used for each of the three grade level datasets, but for each grade level all test forms were calibrated simultaneously. This resulted in all items being calibrated on the same scale within a given grade level. Most of the Parscale defaults values were used during the calibration; however a log-normal prior distribution was used for estimating the slope parameters, and a normal prior distribution was used in estimating the threshold parameters. Examinee abilities were estimated using an *expected a posterior* estimation procedure with a N(0,1) prior distribution specified. Both item parameters and person parameters were saved to external files, and the l_z, l_{zh}, and *PM* indices were calculated for each examinee.

Evaluation of Indices. To evaluate the performance of the indices several different methods were used. The first four moments of the distributions of l_z and l_{zh} were determined, and the correlation of these two indices with $\hat{\theta}$ was determined. This was used to determine if these indices followed a standard distribution. Because of the sample size involved the Kolmogorov-Smirnov test[2] was not performed. If the l_z and l_{zh} indices tend to follow a standard normal distribution then future measurement specialists will know what to expected in terms of detection rates within the framework of a statewide assessment program. Having the l_z and l_{zh} follow a standard normal distribution will also allow for simple interpretation. For example, if l_z follows the expected distribution then values greater than 1.96 in absolute value will result in approximately 5% of the examinees being identified as having nonmodel fitting $\hat{\theta}$ values. Similar evaluation of the *PM* approach was not done as its theoretical distribution has not been established.

The rate at which examinees were identified as having responded to test questions in a manner that did not reflect $\hat{\theta}$ was also found. For l_z and l_{zh} both one-tailed and two-tailed tests were performed. In the case of the two-tailed test examinees having high positive fit indices would be considered *hyperconsistent* (i.e., a Guttman like response pattern) and examinees with large negative fit indices would be considered *inconsistent*. An α level of 0.01 and also 0.001 was used to classify examinees as nonmodel fitting under both the one-tailed and two-tailed tests. In the case of the one-tailed test l_z and l_{zh} values less than -2.33 (α=0.01) and -3.09 (α=0.001) were considered

[2] The Kolomogorov-Smirnov test is a statistical test the compares a given distribution to a normal density function. However, because of the large sample sizes used in this investigation, the null hypothesis would almost always be rejected.

as nonmodel fitting. Likewise, in the case of the two-tailed test fit indices in absolute value greater than 2.58 (α=0.01) and 3.30 (α=0.001) were considered nonmodel fitting. To gain a better understanding of the relationship between examinee-model fit and ability level the mean and standard deviation of $\hat{\theta}$ for those examinees identified was also found using the l_z and l_{zh} indices. Using the *PM* approach examinees identified with α=0.01 and α=0.001 were identified, and the mean and standard deviation of $\hat{\theta}$ was also determined for these examinees. The nature of the *PM* approach does not lend itself to a two-tailed test.

A listwise crosstabs analysis was performed where the l_z, l_{zh}, and *PM* methods were compared. Using this method the number of examinees identified with one or more of the methods can be determined. For example, using this method it can be determined what number of students were uniquely identified as nonmodel fitting using the *PM* approach. Using the listwise approach will allow us to determine the number of examinees that are commonly and uniquely identified as nonmodel fitting across the fit indices.

Finally, example response vectors will be presented that will demonstrate the types of response patterns identified using the *PM* method and the l_z and l_{zh} values will also be presented. The response patterns presented are from actual examinees responding to math test questions on the NHEIAP assessment. These examples will allow for us to demonstrate the types of examinee response patterns that are identified using the *PM* method, and to further investigate the relationship among the various person-fit indices.

Results

Distributional Characteristics. The distributional characteristics of the l_z and the l_{zh} indices are presented in Table 1. A cursory review of this table suggests that the distributions of these indices tend to follow a standard distribution across the various grade levels, although the mean scores for l_z were somewhat different from 0 and the standard deviations were for both statistics were smaller than 1 (with the exception of l_z in grade 10). The mean values for the l_{zh} index were all 0.00 while l_z had mean values ranging from 0.14 to -0.24. For both the l_z and l_{zh} indices the standard deviation values all ranged from approximately 0.80 to 1.00. The indices of skewness and kurtosis presented in Table 1 also suggest that the indices are approximately normally distributed. The only exception appears to be the kurtosis for the l_{zh} indices that ranged from 1.66 to 2.20; thus, suggesting a small degree of leptokurtosis in the distributions.

Also presented in Table 1 are the correlation indices between the person-fit indices and $\hat{\theta}$ values. For the l_{zh} indices these values were all 0.00, and for l_z these values ranged from 0.03 to 0.15. (Significance testing was not performed because of sample size.) Overall, the results in Table 1 suggest that both the l_z and l_{zh} indices tend to be distributed in a manner that is expected, and that the indices are not dependent on ability levels. Thus, examinees across the ability continuum have an equal chance of being identified as nonmodel fitting using the l_z and l_{zh} indices.

Detection Rates. In Tables 2 through 4 contain the detection rates of l_z, l_{zh}, and *PM* along with the characteristics associated with the ability of those examinees identified. In Table 2 (grade 3) more examinee were identified as nonmodel fitting using the *PM* method compared to either the l_z or l_{zh} indices. For example, 502

examinees were identified with *PM* (p<0.01) while 178 and 187 examinees were identified as nonmodel fitting using the l_z and l_{zh} indices respectively. This was also observed at grades 6 and 10 as can be seen in Tables 3 and 4, respectively.

Another noticeable difference among the various methods is that the *PM* method appears to identify examinees that are less able than either the l_z or l_{zh} methods. For example, in grade 6 (Table 3) the average $\hat{\theta}$ value for examinees identified with the *PM* method was -1.25 (at the p=0.001 level) while for the l_z and l_{zh} indices this value was -0.62 and -0.70, respectively. This result is not surprising given that motivation was the main principle that lead to the development of *PM*, and that is reasonable less motivated examinees will have $\hat{\theta}$ values that suggest that they are less able. While the l_z and l_{zh} indices were developed to identify a wider variety of behaviors that would result in response patterns that are not in accordance with the underlying test model.

The results in Tables 2 through 4 also suggest that few examinees are responding in a hyperconsistent manner relative to the model. For example, in Table 4 we see that only 17 and 14 10th graders were identified (p=0.01) as nonmodel fitting in the upper tails of the l_z and l_{zh} indices. Across all three grade levels no examinees were identified as hyperconsistent when p=0.001 for both the l_z and l_{zh} indices. This result is expected because hyperconsistency is usually the result of a cheating behavior (Levine & Rubin, 1979), and in an assessment situation where there is no consequence to the student there is little motivation for cheating.

Crosstabs Analysis. The results of the listwise crosstabs analysis are presented in Tables 5 through 7. In each of these tables there are five columns of inclusion codes. These inclusion codes are dichotomous (0/1) and indicate which indices is being used. In the first line of each of these tables is the number of examinees that were not identified with any of the methods (all inclusion codes equal 0). In the second line of each table is presented the number of students that were uniquely identified by the *PM* method (inclusion codes equal: 00001). When the inclusion codes equal 00011 then the number of students identified by both the *PM* method and the l_{zh} (two tailed) method are presented, and so forth. Only when the frequency is larger than zero for a given set of inclusion codes are the results presented in the listwise crosstabs tables.

Comparing the l_z and l_{zh} indices in Grade 10 (Table 7) we see that these indices appear to identify different students. For example, for α=0.01 and a one-tailed test was used (inclusion codes=10100) we see that only 3 students were identified commonly between the two indices. Similar results were found in grades 3 and 6, and this suggests that that l_z and l_{zh} are operating differently from one another.

The most interesting finding presented in the listwise crosstabs is that the majority of the students identified by the *PM* method are not identified by the l_z or l_{zh} methods. For example, in grade three 437 examinees are identified uniquely by *PM* out of the 502 that were identified by *PM* (see Table 2). Thus, in grade 3 approximately 87% of the students identified by *PM* were uniquely identified with that method. Similar results were found in grades 6 and 10 where approximately 88% and 81%, respectively, were uniquely identified by *PM*.

Example Response Patterns. For each grade level five examinee response patterns were selected that represent the types of repeating patterns that the *PM* method is designed to detect. All the response patterns presented in Table 8 were detected by the *PM* method as being suspicious where $\alpha = 0.001$. The results presented in Table 8 are the most compelling results of this investigation suggesting that the new method is a useful tool in detecting suspicious response patterns. Along with the presentation of the response patterns are examinee $\hat{\theta}$ values, l_z, and l_{z^h} for that examinee. An attempt was made to select examinees that were identified as nonmodel fitting according to *PM* only and also according to the l_z and l_{z^h} indices.

In grade 3 the five response patterns presented are clearly suspicious in nature. The first examinee presented at this grade level responded '4' (i.e., the D response option) to all but one item. According to the l_z (-0.56) and l_{z^h} (-1.13) this examinee is responding in an expected manner given the $\hat{\theta}$ value of -1.85. It is likely that this examinee did not read the questions, and did not take the test seriously in any way. Thus, $\hat{\theta}$ of -1.85 more than likely does not reflect this examinee's θ value. For examinee 2 at grade 3 we see all 1s for items 2 through 8 and all 1s for the last seven items. Again, it is reasonable to consider that this examinee is not taking the test seriously for a significant portion of the test; however, according to l_z and l_{z^h} this examinee was responding in a manner that was expected according to his/her ability level. Examinee 5 appears to have a repeating pattern of 2s to many items, and was identified as non-model fitting with both l_z and l_{z^h}.

Similar results were observed in grade 6 and 10. For example, in grade 10 examinee 1 responded 12341234.... through half the test, and then reversed direction with 43214321 for the second half of the test, but fits the model just fine according to l_z and l_{z^h}. Examinees 3 and 4 in grade 10 (these examinees took different forms of the test) responded all 1s and were only identified by l_{z^h} as nonmodel fitting with l_z and l_{z^h} values of –3.30 and –3.36, respectively

Discussion

Within the framework of large-scale assessment programs there is often a situation where there are no or little consequences associated with how an examinee performs. Because of this it is likely that the examinee is performing according to θ, and is not taking the test seriously. In this study we developed a new approach that is designed to identify students that have responded to test question in a manner that does not accurately reflect $\hat{\theta}$. The new method was developed to detect examinees that responded according to some repeating pattern, and the new method was compared to the previously developed l_z and l_{z^h} indices developed by Drasgow, Levine & Williams (1985).

The l_z and l_{z^h} indices were found to be distributed in a manner that was more or less expected (following a standard distribution). This finding is important and suggests that these indices may be useful in detecting examinees that are responding not in accordance with the underlying test model. Although these indices may be useful in detecting certain types of nonmodel fitting behaviors, they were inconsistent in detecting examinees that responded according to a repeating pattern. As presented in Table 8, the results suggest that l_z and l_{z^h} will only occasionally detect this sort of responding behavior. However, given that l_z and l_{z^h} are designed to detect not a

BEST COPY AVAILABLE

specific form of nonmodel-fitting behaviors, these indices may be useful within the framework of large scale assessments. Additional research is needed to further explore how these indices can be used in a meaningful way within this assessment situation.

In using the *PM* method a Bonferonni correction procedure was used to control the probability of making a type I error. There are many other correction procedures that could have been used (Kirk, 1982), and additional research is needed to further explore the various options. Researchers should also consider controlling the rate of false rejection by making use of such methods as the false discovery rate procedure (FDR; Benjamini & Hochberg, 1994). The FDR procedure controls the probability of making even one false rejection in the set of comparisons, and results in a method that controls the expected proportion of falsely rejected hypotheses. Additional research is also needed to determine, possibly through monte carlo simulation, the theoretical distribution of the *PM* index.

The results of this investigation are promising and demonstrate that the *PM* method can be used by measurement specialists working on large-scale assessment programs. Additional research is needed to further explore how this approach can be use in conjunction with other approaches to indexing examinee-model fit. That is, it is reasonable to consider using several different approach in a complimentary manner to determine the extent to which $\hat{\theta}$ is an accurate reflection of θ regardless of the behavior that underlies the responses to test questions. Researchers should consider the use of several person fit methods to be used a profile of person fit.

References

Benjamini, Y., & Hochberg, Y. (1994). Controlling the false discovery rate: Practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, *57*, 289-300.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10*, 167-174.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8:1*.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives, 6:1*.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of American Statistical Association, 56*, 52-64.

Freund , D. S. & Rock, D.A. (1992). A preliminary investigation of pattern-marking in 1990 NAEP data. Paper presented at the annual meeting of the American Educational Research association, San Francisco. (ERIC Document Reproduction Service No. ED 347 189)

Haldyna, T.M. (1994). *Developing and validating multiple-choice test items*. Lawrence Erlbaum, NJ.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56*, 213-228.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*, 215-231.

Meijer, R. R., & Sijtsma, K. (1999). A review of methods for evaluating the fit of item score patterns on a test (Research Report 99-01). Enschede, The Netherlands: University of Twente, Department of Education.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22*, 53-69.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121-129.

Paris, S.G., Lawton, T.A., Turner, J.C., & Roth, J.L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher, 20*, 2-7.

Schmitt, N, Chan, D. Sacco, J.M, McFarland, L.A. & Jennings D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41-53.

Sijtsma, K., & Meijer, R. R. (in press). The person response function as a tool in person fit research. *Psychometrika*.

Van Krimpen-Stoop, E. M. L. A., & Meijer, R.R. (2000). Detection of person misfit in computerized adaptive tests with polytomous items. Submitted for publication.

Wolf, L. F. et al. (1995). Consequences of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.

Table 1
Distributional Characteristics and
Correlation with Ability for l_z and l_{zh} Statistics

Grade Level & Statistic	l_z	l_{zh}
Grade 3		
Mean	0.14	0.00
Standard Deviation	0.87	0.82
Skewness	-0.71	-0.71
Kurtosis	0.98	2.19
Correlation with $\hat{\theta}$	0.15	0.00
Grade 6		
Mean	0.18	0.00
Standard Deviation	0.84	0.79
Skewness	-0.72	-0.62
Kurtosis	1.35	2.20
Correlation with $\hat{\theta}$	0.11	0.00
Grade 10		
Mean	-0.24	0.00
Standard Deviation	1.01	0.90
Skewness	-0.46	-0.63
Kurtosis	0.35	1.66
Correlation with $\hat{\theta}$	0.03	0.00

Table 2
Grade 3 (N=16,630)
Mean and Standard Deviation of
Ability Estimates for Students Identified with l_z, l_{zh}, and
the PM Indices Under Type I Error Rates of $p<0.01$, and $p<0.001$

	$p<0.01$			$p<0.001$		
Index	Mean	St Dev	N	Mean	St Dev	N
One tailed						
l_z	-0.71	0.54	178	-0.77	0.50	34
l_{zh}	-0.80	0.70	187	-1.04	0.67	38
Two tailed						
Upper tail						
l_z	-0.95	0.18	3	n/a	n/a	0
l_{zh}	-1.62	0.13	4	n/a	n/a	0
Lower tail						
l_z	-0.68	0.57	102	-0.71	0.53	23
l_{zh}	-0.88	0.72	120	-1.04	0.66	39
PM	-1.16	0.44	502	-1.32	0.39	118

Table 3
Grade 6 (N=16,387)
Mean and Standard Deviation of
Ability Estimates for Students Identified with l_z, l_{zh}, and
the PM Indices Under Type I Error Rates of $p<0.01$, and $p<0.001$

	$p<0.01$			$p<0.001$		
Index	Mean	St Dev	N	Mean	St Dev	N
One tailed						
l_z	-0.56	0.43	148	-0.62	0.40	34
l_{zh}	-0.80	0.69	149	-0.70	1.03	48
Two tailed						
Upper tail						
l_z	-0.57	0.15	3	n/a	n/a	0
l_{zh}	-1.31	0.25	7	n/a	n/a	0
Lower tail						
l_z	-0.57	0.40	85	-0.64	0.36	23
l_{zh}	-0.63	1.12	39	-0.78	0.78	99
PM	-1.11	0.50	362	-1.25	0.48	65

Table 4
Grade 10 (N=13,445)
Mean and Standard Deviation of
Ability Estimates for Students Identified with l_z, l_{zh}, and
the PM Indices Under Type I Error Rates of $p<0.01$, and $p<0.001$

	$p<0.01$			$p<0.001$		
Index	Mean	St Dev	N	Mean	St Dev	N
One tailed						
l_z	0.12	0.48	404	0.10	0.41	87
l_{zh}	-0.54	0.51	211	-0.60	0.22	69
Two tailed						
Upper tail						
l_z	-0.12	0.28	17	n/a	n/a	0
l_{zh}	-0.82	0.22	14	n/a	n/a	0
Lower tail						
l_z	0.15	0.46	247	0.05	0.35	62
l_{zh}	-0.57	0.52	166	-0.62	0.21	53
PM	-0.76	0.41	257	-0.86	0.35	62

Table 5
Crosstabs Analysis for Grade 3
(OT=one tailed, TT=two tailed)

Inclusion Codes						
l_z OT	l_z TT	l_{z^b} OT	l_{z^b} TT	PM	Frequency	Percentage
α=0.01						
0	0	0	0	0	15875	95.5
0	0	0	0	1	437	2.6
0	0	0	1	0	4	0.0
0	0	1	0	0	45	0.3
0	0	1	0	1	10	0.1
0	0	1	1	0	48	0.3
0	0	1	1	1	30	0.2
0	1	0	0	0	3	0.0
1	0	0	0	0	56	0.3
1	0	0	0	1	3	0.0
1	0	1	0	0	3	0.0
1	0	1	0	1	1	0.0
1	0	1	1	0	10	0.1
1	0	1	1	1	3	0.0
1	1	0	0	0	57	0.3
1	1	0	0	1	8	0.0
1	1	1	0	0	8	0.0
1	1	1	1	0	19	0.1
1	1	1	1	1	10	0.1
α=0.001						
0	0	0	0	0	16443	98.9
0	0	0	0	1	112	0.7
0	0	1	0	0	11	0.1
0	0	1	1	0	28	0.2
0	0	1	1	1	2	0.0
1	0	0	0	0	6	0.0
1	0	0	0	1	1	0.0
1	0	1	0	0	1	0.0
1	0	1	1	0	3	0.0
1	1	0	0	0	15	0.1
1	1	0	0	1	1	0.0
1	1	1	0	0	1	0.0
1	1	1	1	0	4	0.0
1	1	1	1	1	2	0.0

Table 6
Crosstabs Analysis for Grade 6
(OT=one tailed, TT=two tailed)

Inclusion Codes						
l_z OT	l_z TT	l_{zh} OT	l_{zh} TT	PM	Frequency	Percentage
α=0.01						
0	0	0	0	0	15804	96.4
0	0	0	0	1	317	1.9
0	0	0	1	0	7	0.0
0	0	1	0	0	34	0.2
0	0	1	0	1	7	0.0
0	0	1	1	0	52	0.3
0	0	1	1	1	15	0.1
0	1	0	0	0	3	0.0
1	0	0	0	0	49	0.3
1	0	0	0	1	5	0.0
1	0	1	0	0	2	0.0
1	0	1	1	0	6	0.0
1	0	1	1	1	1	0.0
1	1	0	0	0	45	0.3
1	1	0	0	1	6	0.0
1	1	1	0	0	6	0.0
1	1	1	0	1	1	0.0
1	1	1	1	0	17	0.1
1	1	1	1	1	10	0.1
α=0.001						
0	0	0	0	0	16250	99.2
0	0	0	0	1	62	0.4
0	0	1	0	0	9	0.1
0	0	1	1	0	28	0.2
0	0	1	1	1	2	0.0
1	0	0	0	0	9	0.1
1	0	1	1	0	2	0.0
1	1	0	0	0	15	0.1
1	1	0	0	1	1	0.0
1	1	1	1	0	7	0.0

Table 7
Crosstabs Analysis for Grade 10
(OT=one tailed, TT=two tailed)

Inclusion Codes						
l_z OT	l_z TT	l_{zh} OT	l_{zh} TT	PM	Frequency	Percentage
α=0.01						
0	0	0	0	0	12650	94.1
0	0	0	0	1	207	1.5
0	0	0	1	0	13	0.1
0	0	0	1	1	1	0.0
0	0	1	0	0	31	0.2
0	0	1	0	1	3	0.0
0	0	1	1	0	103	0.8
0	0	1	1	1	34	0.3
0	1	0	0	0	1	0.0
1	0	0	0	0	138	1.0
1	0	0	0	1	3	0.0
1	0	1	0	0	3	0.0
1	0	1	0	1	2	0.0
1	0	1	1	0	10	0.1
1	0	1	1	1	1	0.0
1	1	0	0	0	218	1.6
1	1	0	0	1	3	0.0
1	1	1	0	0	5	0.0
1	1	1	0	1	1	0.0
1	1	1	1	0	18	0.1
1	1	1	1	1	2	0.0
α=0.001						
0	0	0	0	0	13239	98.5
0	0	0	0	1	51	0.4
0	0	1	0	0	13	0.1
0	0	1	0	1	3	0.0
0	0	1	1	0	45	0.3
0	0	1	1	1	7	0.1
0	0	1	1	0	1	0.0
0	0	1	1	1	1	0.0
1	0	0	0	0	24	0.2
1	0	0	0	1	1	0.0
1	1	0	0	0	60	0.4
1	1	1	0	0	1	0.0
1	1	1	1	0	1	0.0

Table 8
Example Students Response Patterns for Examinees
Identified as Having Suspicious Response Patterns Using the *PM* Method

Student	$\hat{\theta}$	l_z	l_{zh}	Response Pattern
Grade 3				
1	-1.85	-0.56	-1.13	44444444444414444444444444444444444444
2	-1.25	-1.32	-0.65	41111111221114211212231234123141311111111
3	-1.46	-1.42	-2.13	33111111121114211212333333333333333333
4	-1.07	-1.13	-3.03	33231323333333313223123333344213443433 23
5	-1.14	-3.67	-3.80	24222323442232322223231233121423312322 22
Grade 6				
1	-1.77	-0.06	-0.79	13441344441414114411144411411134 14
2	-1.53	0.57	-1.65	43433444441443123444444441122212 43
3	-1.27	-1.52	-2.39	11111111111111111111111111111111111
4	-0.83	-0.50	-3.02	41314433324131214213432411234312 43
5	-0.97	-2.92	-4.93	31412314314231244324312432214142 31
Grade 10				
1	-1.34	0.20	-1.50	12341234123412341432143214321432143 14
2	-0.49	-1.00	-0.13	33333333433333333133331322341234 41
3	-0.75	-1.23	-3.14	11111111111111111111111111111111111
4	-0.76	-1.30	-3.30	11111111111111111111111111111111111
5	-0.79	-2.35	-3.68	12344321123443212214334122 14 34124

BEST COPY AVAILABLE